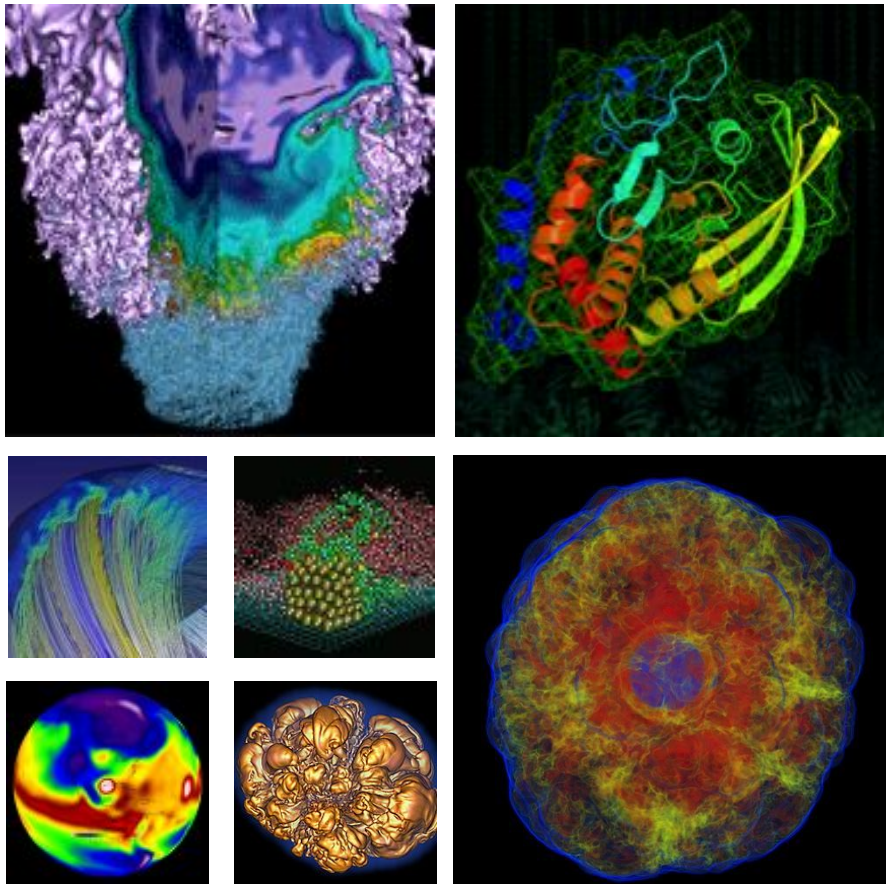
Data Visualization at NERSC



Annette Greiner, NERSC Data and Analytics Services

Why Visualize?

Anscombe's Quartet

X	Y
10.0	8.04
8.0	6.95
13.0	7.58
9.0	8.81
11.0	8.33
14.0	9.96
6.0	7.24
4.0	4.26
12.0	10.84
7.0	4.82
5.0	5.68

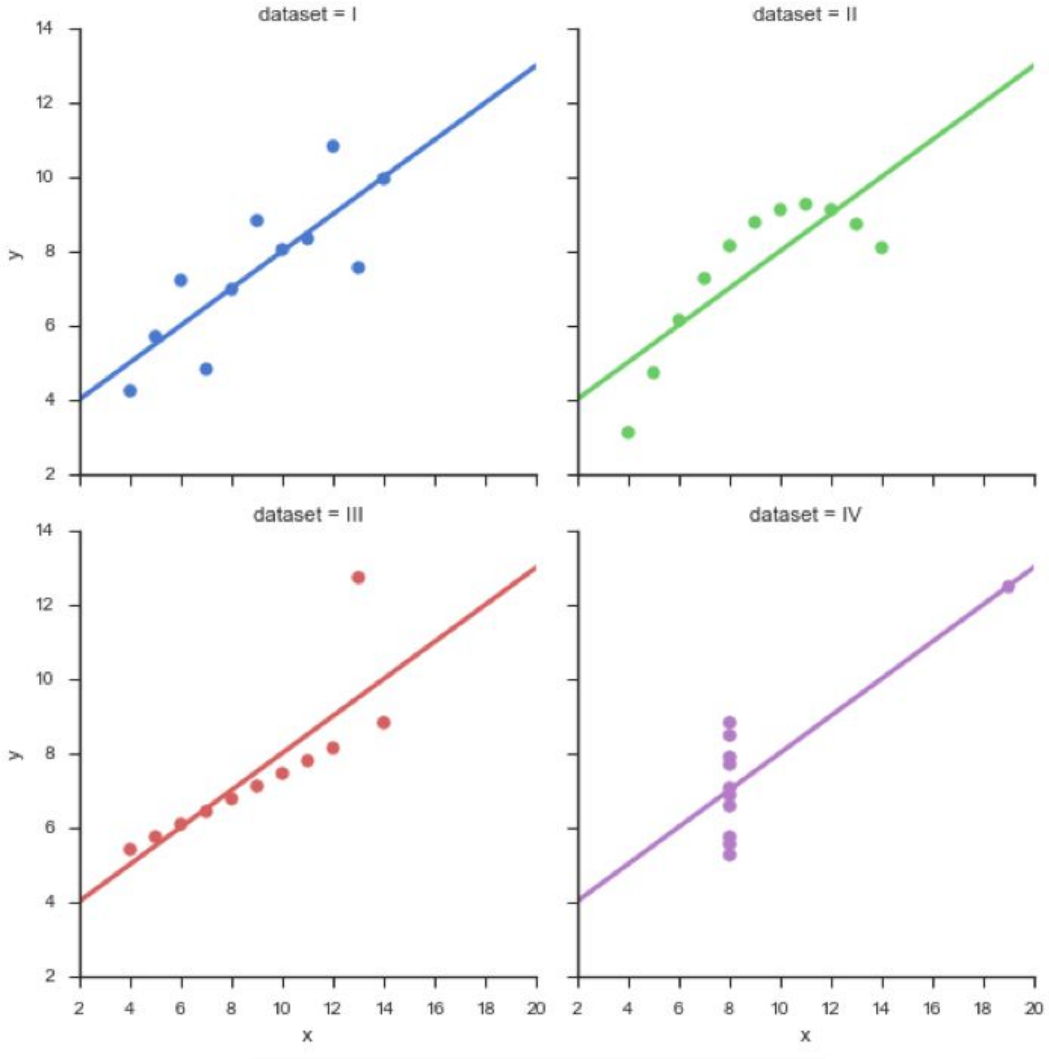
X	Y
10.0	9.14
8.0	8.14
13.0	8.74
9.0	8.77
11.0	9.26
14.0	8.10
6.0	6.13
4.0	3.10
12.0	9.13
7.0	7.26
5.0	4.74

X	Y
10.0	7.46
8.0	6.77
13.0	12.74
9.0	7.11
11.0	7.81
14.0	8.84
6.0	6.08
4.0	5.39
12.0	8.15
7.0	6.42
5.0	5.73

X	Y
8.0	6.58
8.0	5.76
8.0	7.71
8.0	8.84
8.0	8.47
8.0	7.04
8.0	5.25
19.0	12.50
8.0	5.56
8.0	7.91
8.0	6.89

<u>Same:</u>
N
Mean X
Mean Y
Variance X
Variance Y
Regression
Correlation

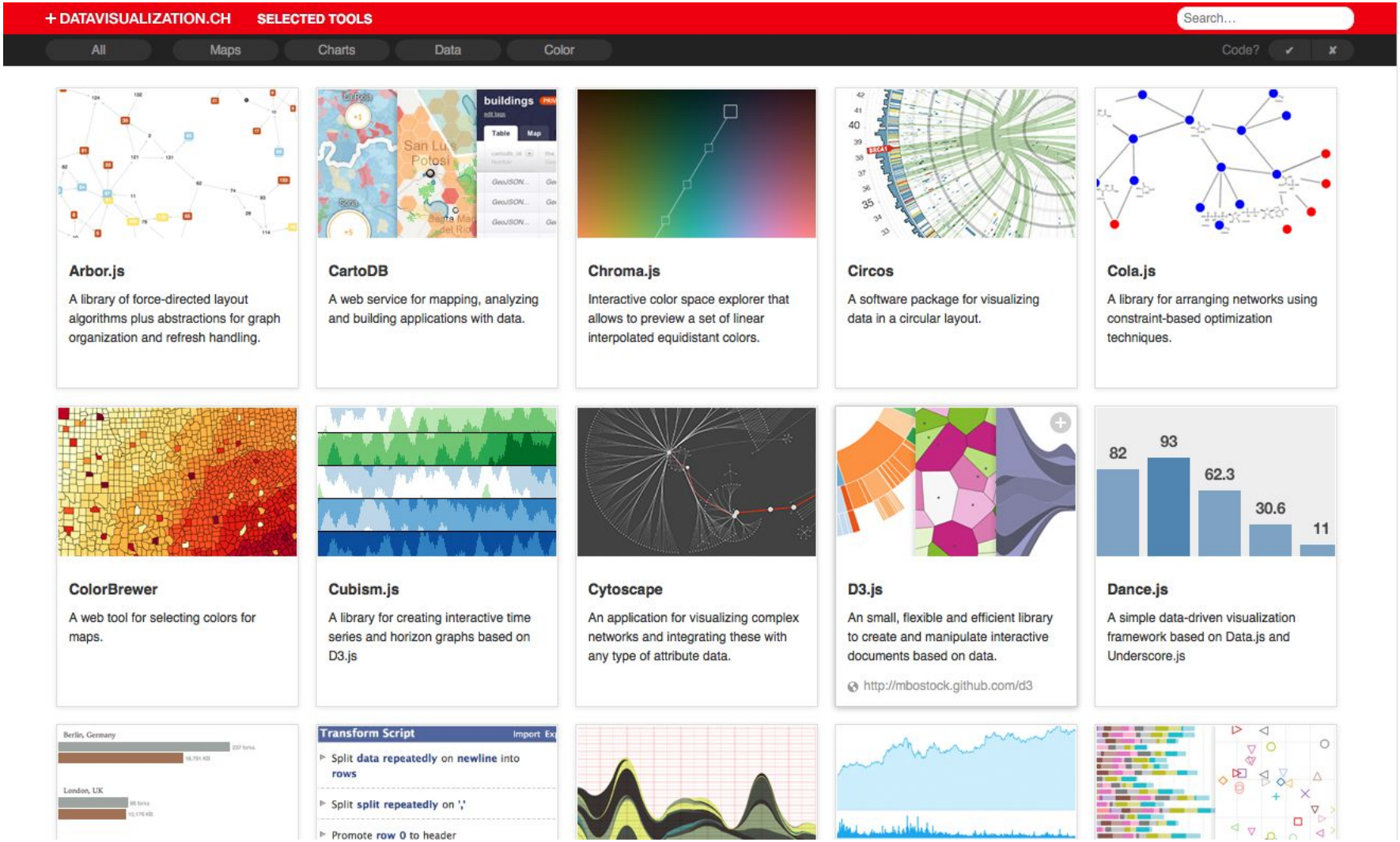
Anscombe's Quartet



F.J. Anscombe, 1973

Choosing a Tool

Good News/Bad News



Goal

Is your goal explanation or exploration?

Explanatory:	Exploratory:
You already know what you want to say.	You want to find out what the data means.

Speed/Flexibility Tradeoff

Do you have lots of time? Do you need customization?

Speed, Ease

Flexibility

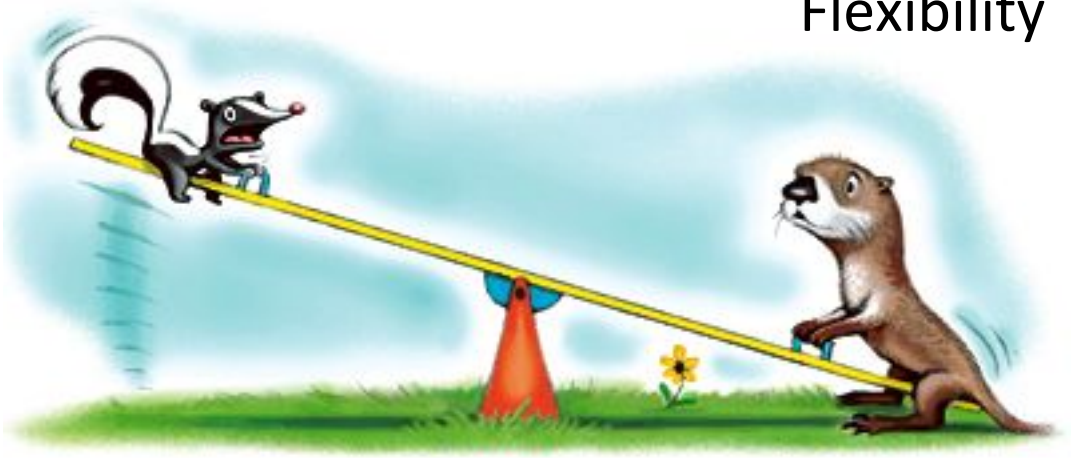
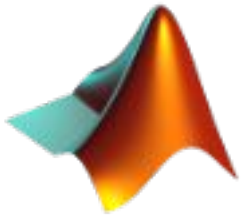


Image: Highlights for Children

Dependencies

What is compatible with your working habits?



Python

R

Analytics app

Web

Libraries

Versions

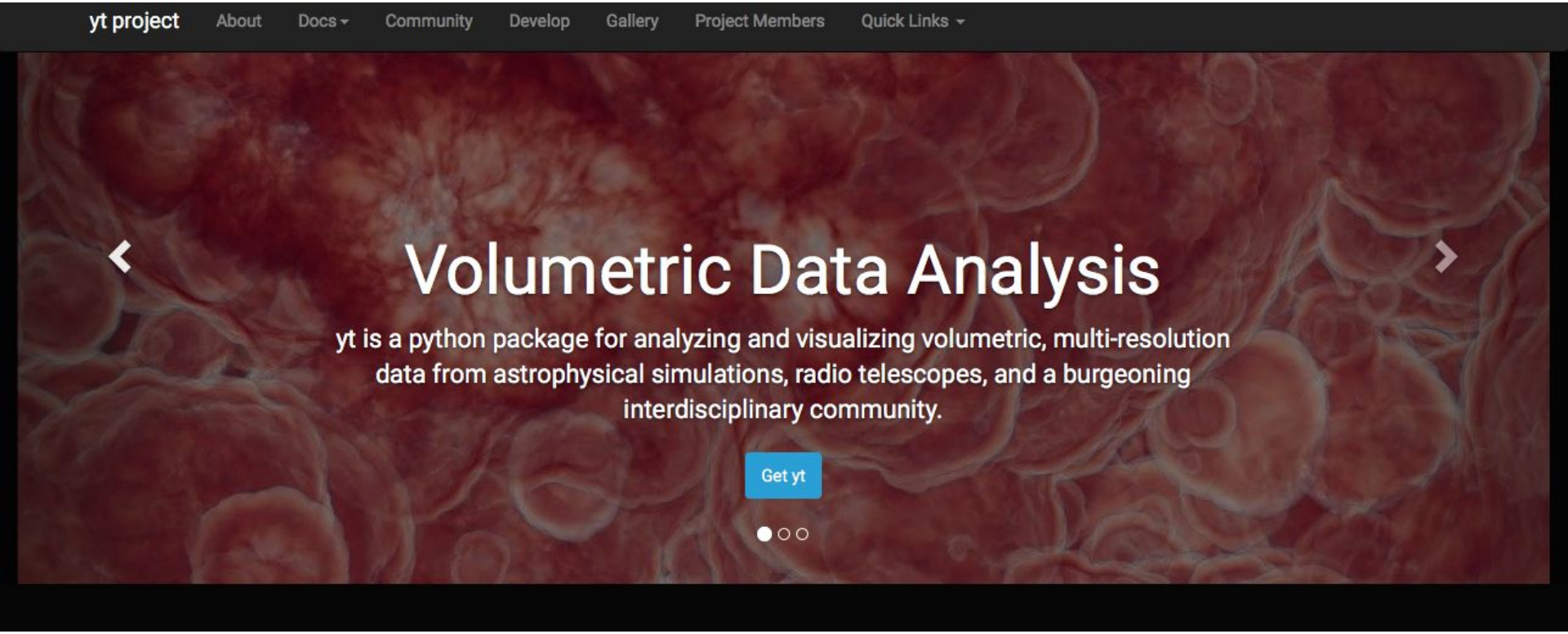
Sharing

Do you need to share with others?

- **Image files (PNG, TIFF, PDF, etc.)**
- **Interactivity (click, brush, rollover, zoom, pan, etc.)**
- **Code sharing (notebook interface)**
- **Privacy (authentication, authorization)**

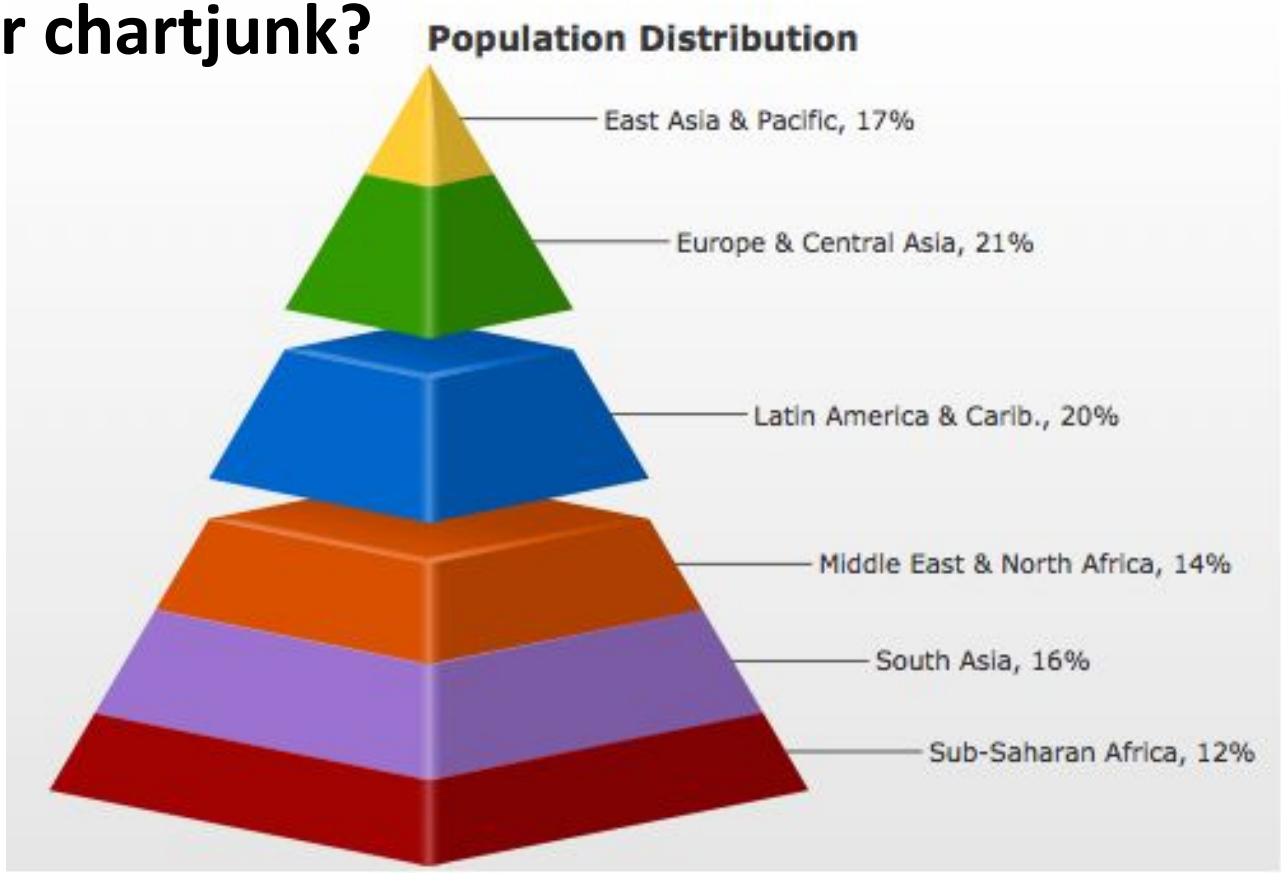
Specificity

Is there a tool aimed at your problem in particular?



Graphical Quality

What tool will let you show your data without distortion or chartjunk?



FusionCharts

Licensing/Cost

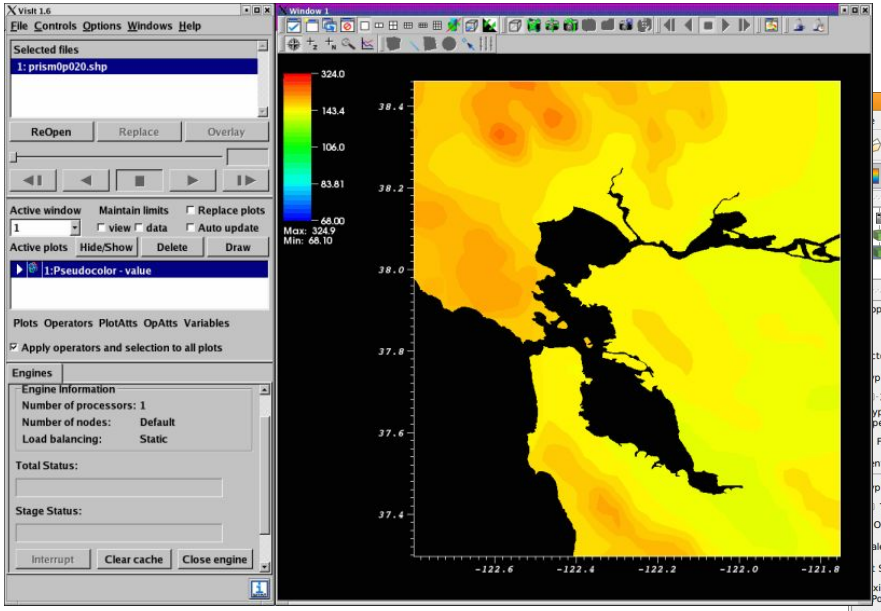
What are the costs?

- **Purchase price**
- **License over time**
- **License constraints (e.g., attribution, constraints on derivative works)**

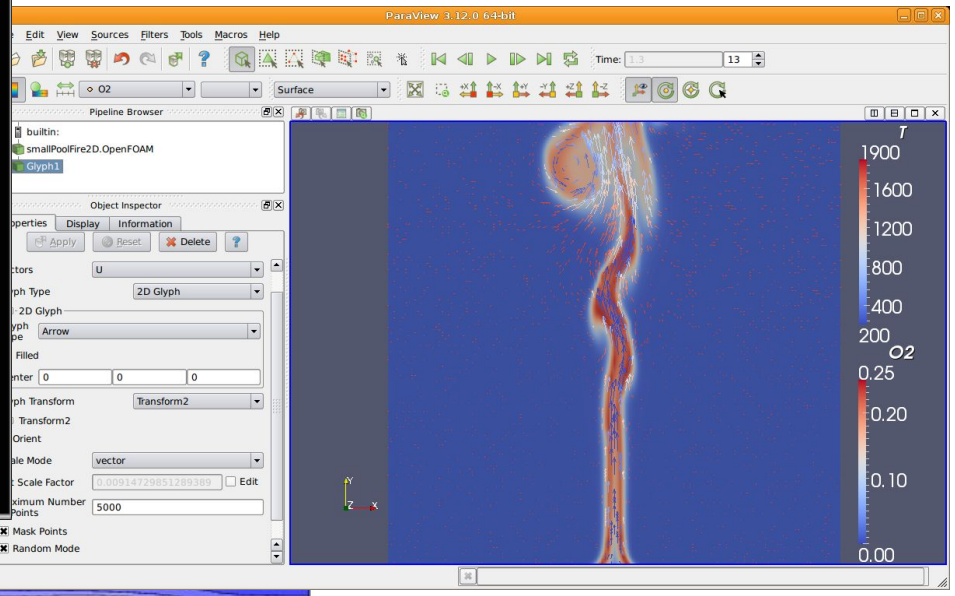
Common Tools at NERSC

Scientific Visualization

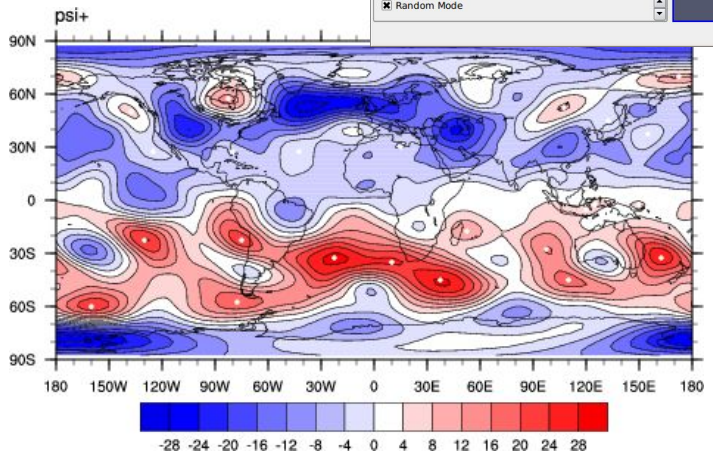
VisIt



ParaView



NCAR Graphics Library

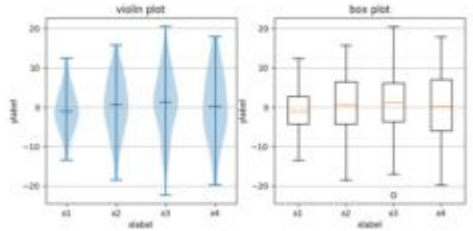


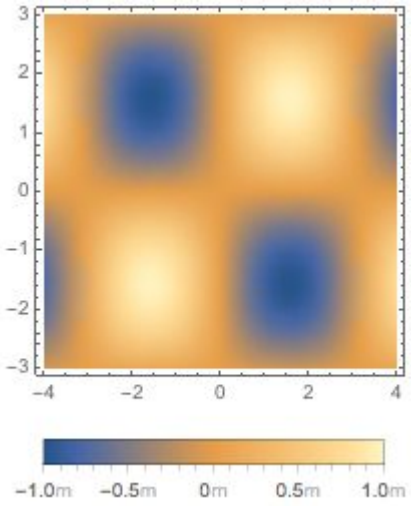
U.S. DEPARTMENT OF ENERGY | Office of Science

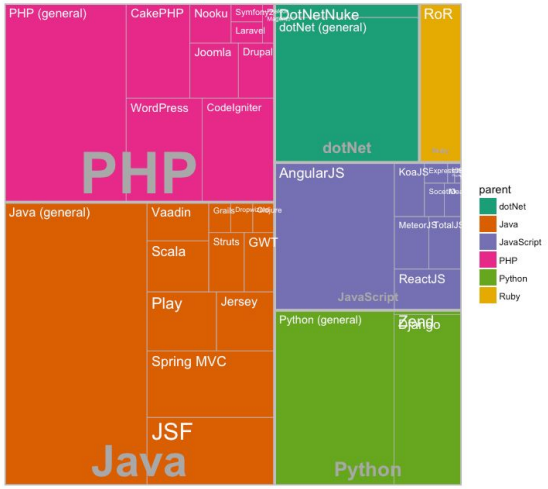
BERKELEY LAB

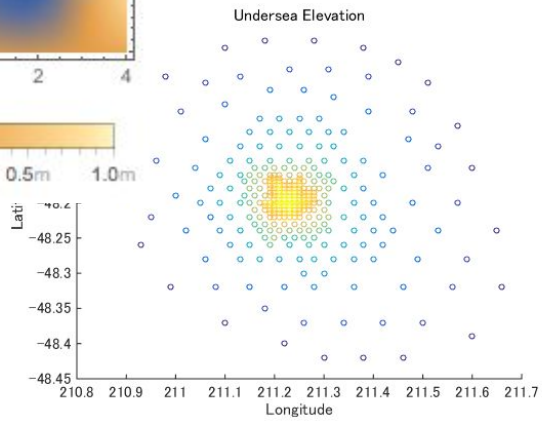
Information Visualization

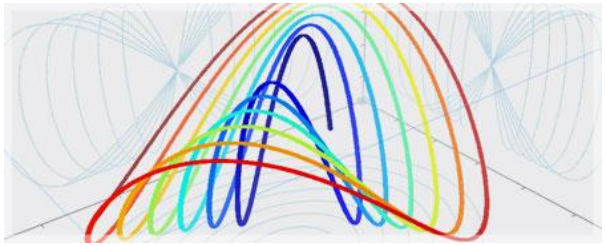
- Libraries for Python (e.g., Matplotlib)
- Libraries for R (e.g., ggplot2)
- Mathematica
- Matlab
- IDL





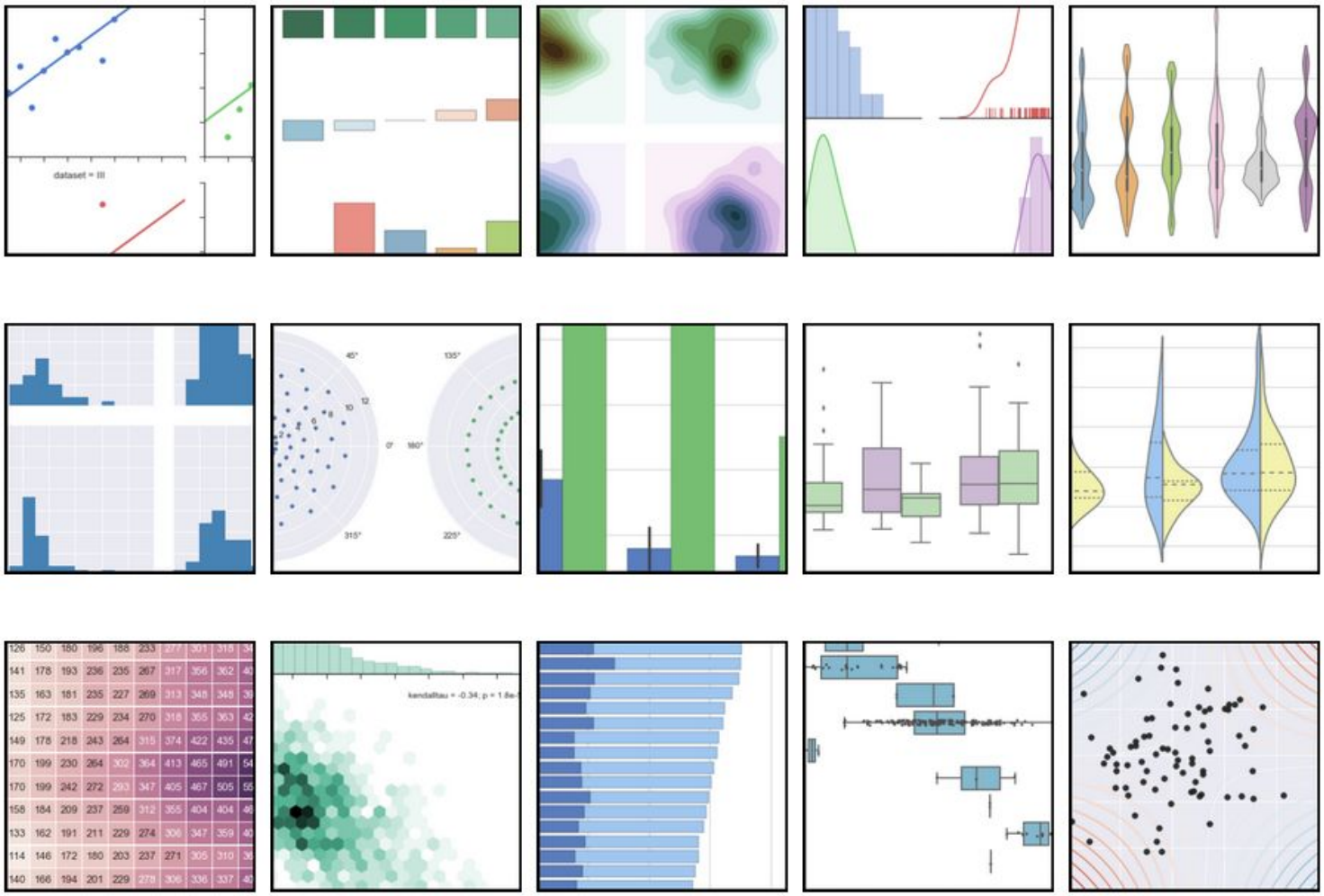






New Tools of Note

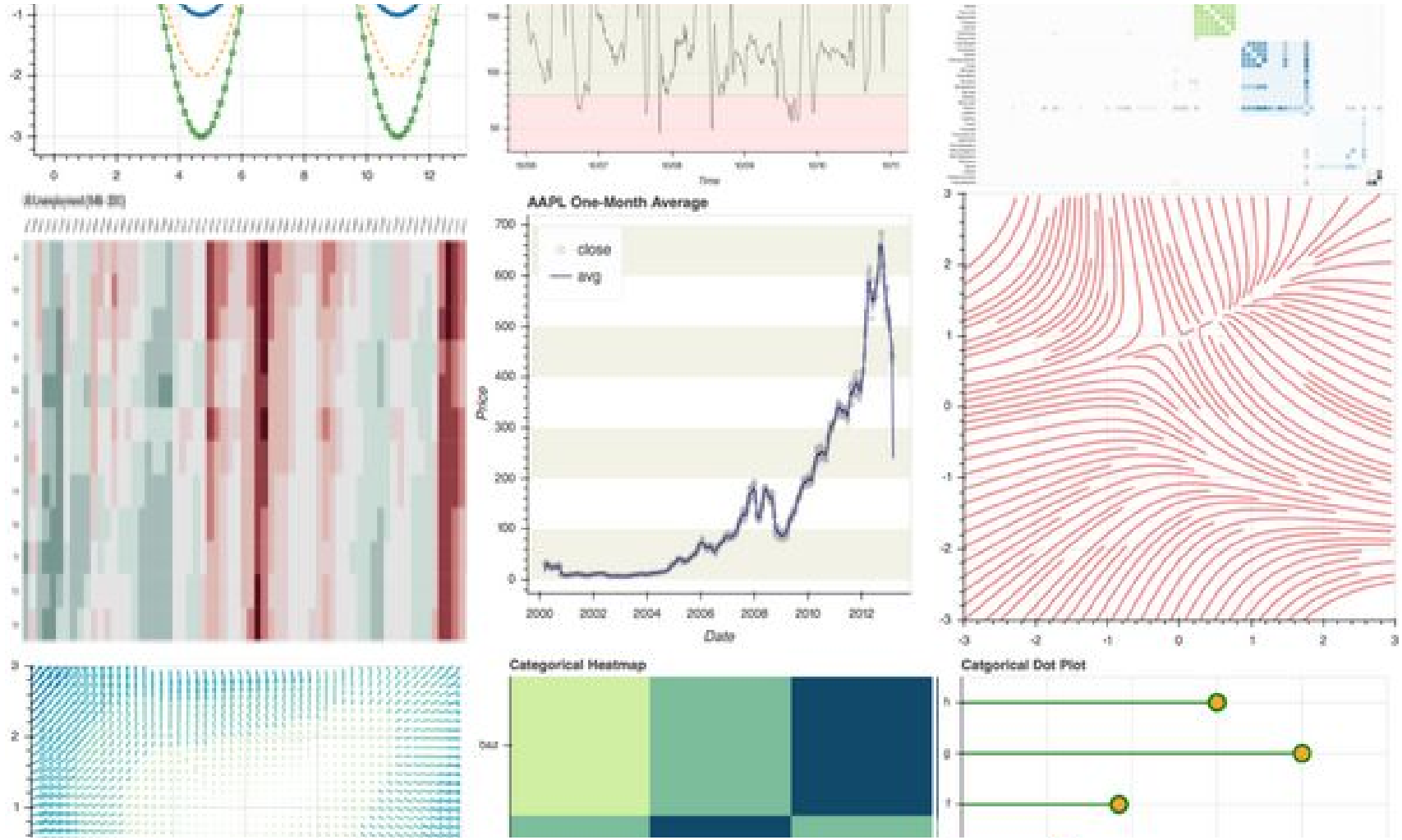
Seaborn



Seaborn

- Goal: exploratory
- Speed/Flexibility: Speed
- Dependencies: python 2.7 or 3.3+, numpy, scipy, matplotlib, pandas
- Sharing: export image files
- Specific Uses: statistical graphs
- Graphical Quality: good quality graphs, ability to tweak colors, axes, etc.
- Costs: open source

Bokeh

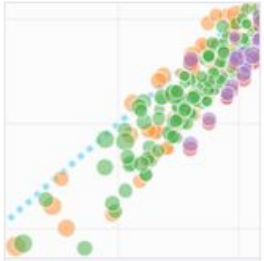

AAPL One-Month Average
close
avg
Price
Date
Categorical Heatmap
Catgorical Dot Plot

Bokeh

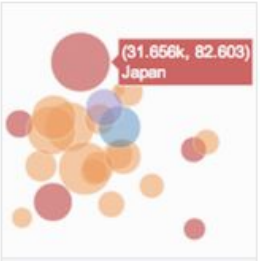
- Goal: exploratory or explanatory
- Speed/Flexibility: low-, intermediate-, or high-level
- Dependencies: python
- Sharing: Bokeh server, embedding in web pages or notebooks
- Specific Uses: novel interactive visualizations in the browser
- Graphical Quality: very good quality
- Costs: open source

Plotly

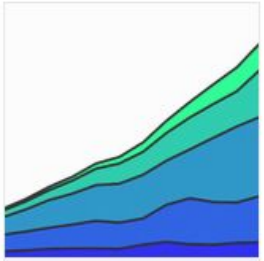
Dashboards



Line and Scatter Plots

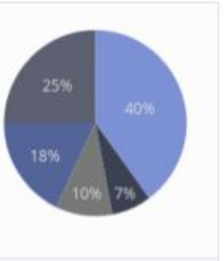


Bubble Charts

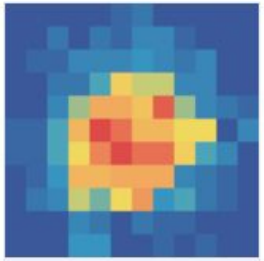


Filled Area Plots

Bar Charts



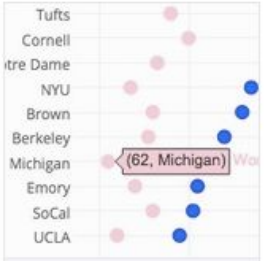
Pie Charts



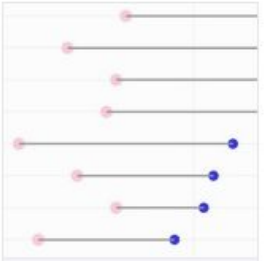
2D Histograms

Range Sliders and Selectors

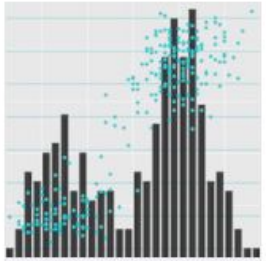
Gauge Chart



Dot plots

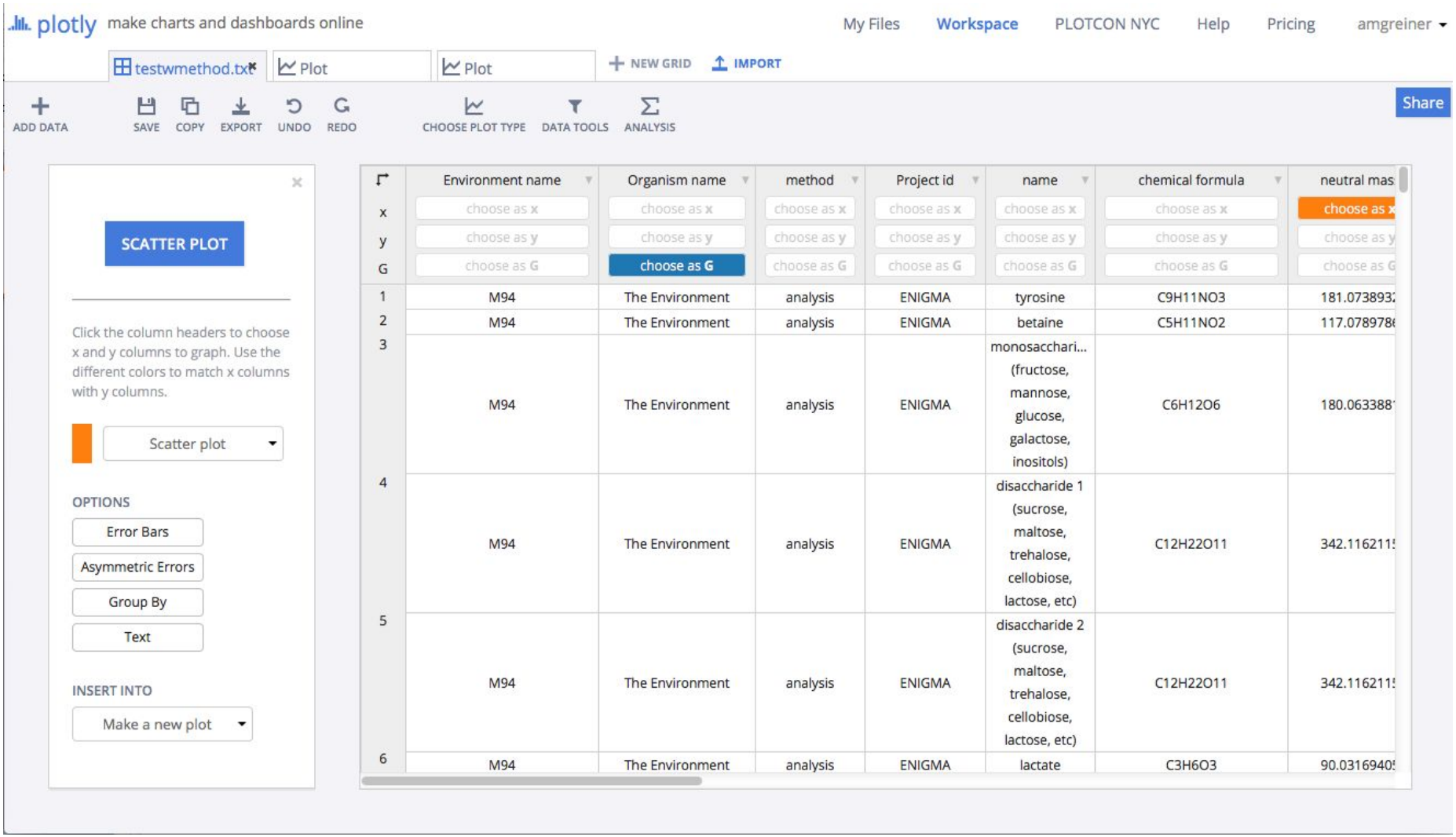


Dumbbell plots



Graphing Multiple Chart Types

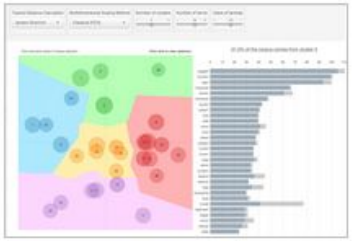
Plotly


plotly make charts and dashboards online
My Files
Workspace
PLOTCON NYC
Help
Pricing
amgreiner
testwmethod.txt
Plot
Plot
NEW GRID
IMPORT
ADD DATA
SAVE
COPY
EXPORT
UNDO
REDO
CHOOSE PLOT TYPE
DATA TOOLS
ANALYSIS
Share
SCATTER PLOT
Click the column headers to choose x and y columns to graph. Use the different colors to match x columns with y columns.
Scatter plot
OPTIONS
Error Bars
Asymmetric Errors
Group By
Text
INSERT INTO
Make a new plot
Environment name
Organism name
method
Project id
name
chemical formula
neutral mas
choose as x
choose as y
choose as G
1 M94 The Environment analysis ENIGMA tyrosine C9H11NO3 181.073893
2 M94 The Environment analysis ENIGMA betaine C5H11NO2 117.078978
3 M94 The Environment analysis ENIGMA monosaccharides (fructose, mannose, glucose, galactose, inositols) C6H12O6 180.063388
4 M94 The Environment analysis ENIGMA disaccharide 1 (sucrose, maltose, trehalose, cellobiose, lactose, etc) C12H22O11 342.116211
5 M94 The Environment analysis ENIGMA disaccharide 2 (sucrose, maltose, trehalose, cellobiose, lactose, etc) C12H22O11 342.116211
6 M94 The Environment analysis ENIGMA lactate C3H6O3 90.031694

Plotly

- Goal: exploratory
- Speed/Flexibility: Speed
- Dependencies: library for python, R, Matlab, Excel, or JS, or use web app
- Sharing: Plot.ly, Plotly Server
- Specific Uses: collaborative analysis
- Graphical Quality: good quality
- Costs: plotly libs are open source, Plotly Server is licensed

Shiny



MULTIDIMENSIONAL SCALING

Similarity analysis tool.

INTERACTIVE REPORTING

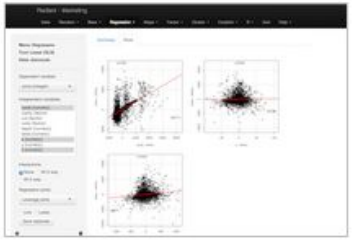
Interactive report with Shiny and R Markdown

GEOMETRY OF CLASSIFIERS

Comparison of machine learning algorithms.

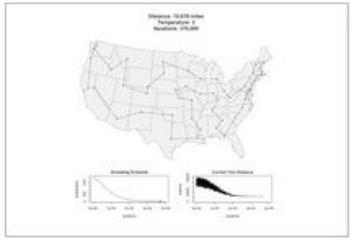
NOMOGRAM GENERATOR

Tune analysis then export as pdf, HTML, or Word file.



RADIANT

Extensive app for teaching business analytics. (documentation)

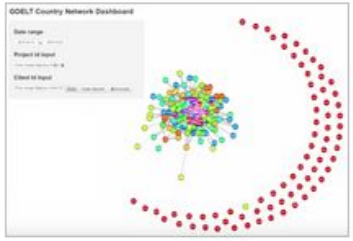


TRAVELING SALESMAN

Optimization fun.

FILE CONVERTER

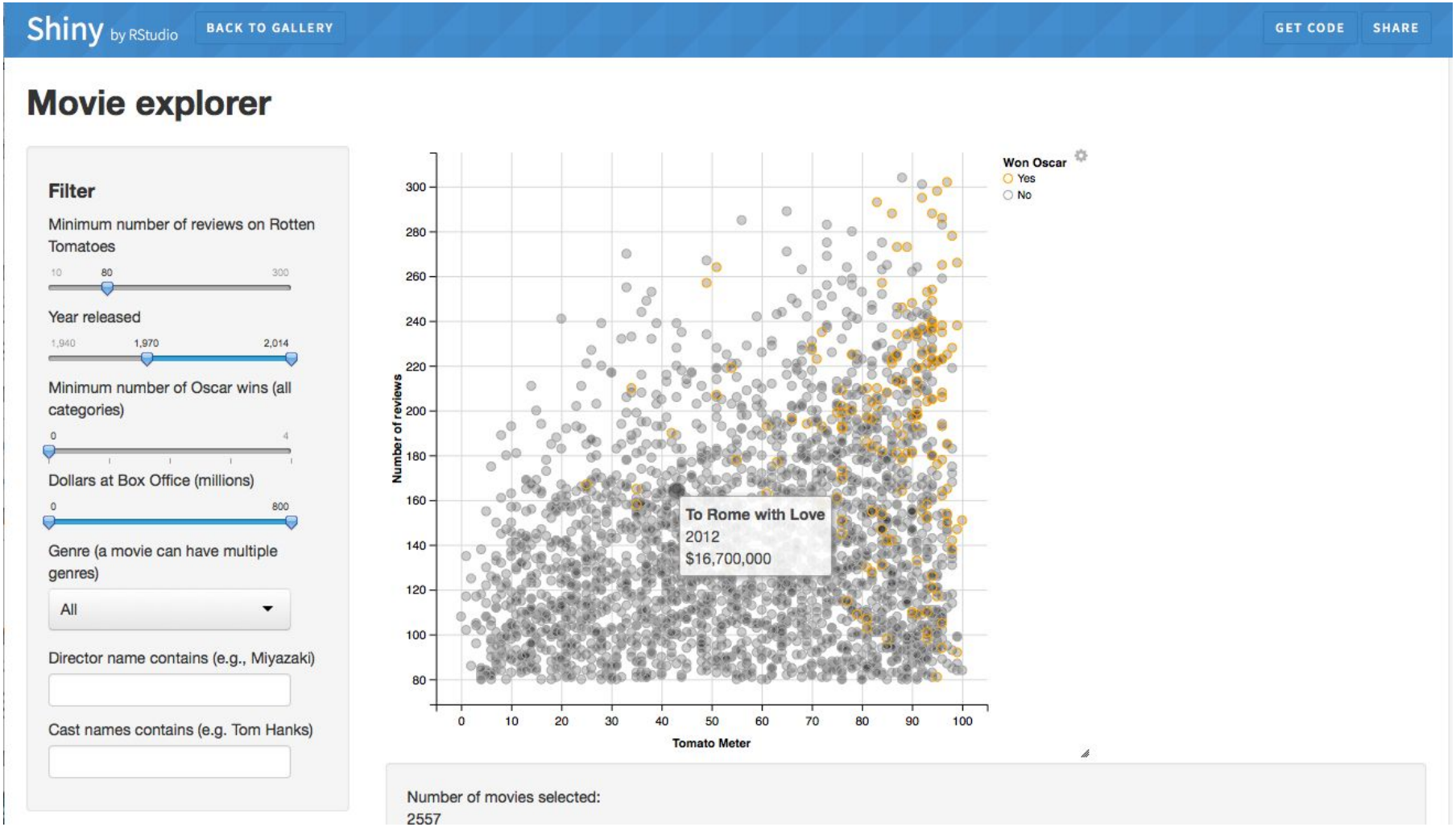
Upload a data file, then download in various formats.



GDELT & BIGQUERY

Dashboard to GDELT database with quarter-billion records

Shiny


Shiny by RStudio
BACK TO GALLERY
GET CODE
SHARE
Movie explorer
Filter
Minimum number of reviews on Rotten Tomatoes
10
80
300
Year released
1,940
1,970
2,014
Minimum number of Oscar wins (all categories)
0
4
Dollars at Box Office (millions)
0
800
Genre (a movie can have multiple genres)
All
Director name contains (e.g., Miyazaki)
Cast names contains (e.g. Tom Hanks)
Won Oscar
Yes
No
Number of reviews
300
280
260
240
220
200
180
160
140
120
100
80
0
10
20
30
40
50
60
70
80
90
100
Tomato Meter
To Rome with Love
2012
$16,700,000
Number of movies selected:
2557

Shiny

- Goal: exploratory or explanatory
- Speed/Flexibility: Speed
- Dependencies: R or R Studio
- Sharing: shinyapps.io, Shiny Server
- Specific Uses: interactive web applications
- Graphical Quality: good quality
- Costs: open source, freemium model

Questions?

Questions?

I have some for you . . .

Questions?

- **What tools do you guys use (on or off HPC)?**

Questions?

- What tools do you guys use (on or off HPC)?
- What would you like to see available at NERSC?

Questions?

- What tools do you guys use (on or off HPC)?
- What would you like to see available at NERSC?
- Do you currently do data vis on HPC systems?

Questions?

- What tools do you guys use (on or off HPC)?
- What would you like to see available at NERSC?
- Do you currently do data vis on HPC systems?
- If no, why not? Are there roadblocks we can remove?

National Energy Research Scientific Computing Center